Varying Receptive Fields for Thermal Image Super-Resolution

Nolan B. Gutierrez and William J. Beksi

Department of Computer Science and Engineering, University of Texas at Arlington

The Goal

Determine a mapping between a low-resolution (LR) image an its high-resolution (HR) counterpart in the HR space.

Introduction

Advances in thermographic cameras have spurred research in the thermal image domain for a variety of applications, but the inherent LR of thermal imaging systems necessitates image restoration techniques. Super-resolution (SR) is an image restoration method which involves the discovery of a mapping from low-quality to high-quality images in the infrared image space. Besides applications in the visible spectrum, there are many uses of SR in the infrared (invisible) range including building inspection, military operations, maritime search and rescue, and much more.

Thermal cameras are subject to deteriorating computing conditions as the degradation of LR images increases thus requiring computationally efficient SR techniques. However, SR algorithms such as convolutional enhancements and visual attention often suffer from performance losses. Dilated convolutions can augment deep SR techniques by increasing receptive fields at a non-increasing computational complexity. By parametrizing several convolutional filters' dilation rate, we can sample from feature maps using distinct receptive fields. Furthermore, we can apply second order channel-attention to efficiently and informatively attend to easily-computed global statistics (e.g., covariance matrices) of feature channels.

Background

Assume $F(\boldsymbol{x};\boldsymbol{\theta}) : \boldsymbol{x} \in \mathcal{R}^{H\times W\times c} \rightarrow \boldsymbol{y} \in \mathcal{R}^{sH\times sW\times c}$, where H, W, and c are the height, weight, and number of channels of an input image. $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{\theta}$, and s are the input image, super-resolved image, parameters associated with F, and scale factor marking the increase of resolution, respectively. We use the following objective (mean squared error)

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \frac{1}{NHW} \sum_{\boldsymbol{x}\in\boldsymbol{X}} \sum_{i=1}^{N} (F(\boldsymbol{x};\boldsymbol{\theta})(i) - \boldsymbol{y}(i))^2 \tag{1}$$

to find $\boldsymbol{\theta}$, where N is the number of sample input images in a batch $\boldsymbol{X}$. $F(\boldsymbol{x};\boldsymbol{\theta})(i)$ and $\boldsymbol{y}(i)$ are the ith pixels of the super-resolved and ground-truth images.

Proposed Method

Dilated Convolutions: As illustrated in the different dilation rates module of Fig. 1, we parametrize parallel convolutional filters thereby introducing distinct dilation rates.

Compression Through Dilations:
An *effective receptive field* (ERF) is defined as the area containing any input pixel with at least some impact on a particular output unit.
Compression through dilations (CTD) is the case in which a convolutional filter uses fewer parameters to increase an ERF with dilated convolutions compared to without dilated convolutions.

Second Order Channel-Attention (SOCA):
First, a feature map of dimension $H \times W \times C$ is reshaped into a feature map $\boldsymbol{X}$ of shape $HW \times C$. Second, the covariance matrix is calculated,

$$\boldsymbol{\Sigma} = \boldsymbol{X}\boldsymbol{I}_{\boldsymbol{f}}\boldsymbol{X}^{\top}, \tag{2}$$

where $\boldsymbol{I_f} = \frac{1}{s}(\boldsymbol{I} - \frac{1}{s}\boldsymbol{1})$ and $s = HW$. $\boldsymbol{I}$ and $\boldsymbol{1}$ are the $m \times m$ identity matrix and the matrix of all ones, respectively. Next, the covariance matrix is pre-normalized,

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{\mathrm{tr}(\boldsymbol{\Sigma})}\boldsymbol{\Sigma}, \tag{3}$$

where $\mathrm{tr}(\cdot)$ denotes the matrix trace. Let $\boldsymbol{Y}_0 = \hat{\boldsymbol{\Sigma}}$ and $\boldsymbol{Z}_0 = \boldsymbol{I}$, then $\boldsymbol{Y_n}$ and $\boldsymbol{Z}_n$ are obtained by

$$\boldsymbol{Y}_n = \frac{1}{2}\boldsymbol{Y}_{n-1}(3\boldsymbol{I} - \boldsymbol{Z}_{n-1}\boldsymbol{Y}_{n-1}), \tag{4}$$

$$\boldsymbol{Z}_n = \frac{1}{2}(3\boldsymbol{I} - \boldsymbol{Z}_{n-1}\boldsymbol{Y}_{n-1})\boldsymbol{Z}_{n-1}, \tag{5}$$

with $\boldsymbol{Y_n}$ and $\boldsymbol{Z}_n$ quadratically converging to $\boldsymbol{Y}$ and $\boldsymbol{Y}^{-1}$, respectively.
The final normalized matrix after five iterations of Newton-Schulz is found by compensating the pre-normalization step with

$$\hat{\boldsymbol{Y}} = \sqrt{tr(\boldsymbol{\Sigma})}\boldsymbol{Y}_N. \tag{6}$$

Afterwards, global covariance pooling is applied to obtain a scalar-valued statistic z_i for each channel i,

$$\boldsymbol{z}_i = \frac{1}{C}\sum_{j}^{C} \hat{\boldsymbol{Y}}_{ij}. \tag{7}$$

Architecture

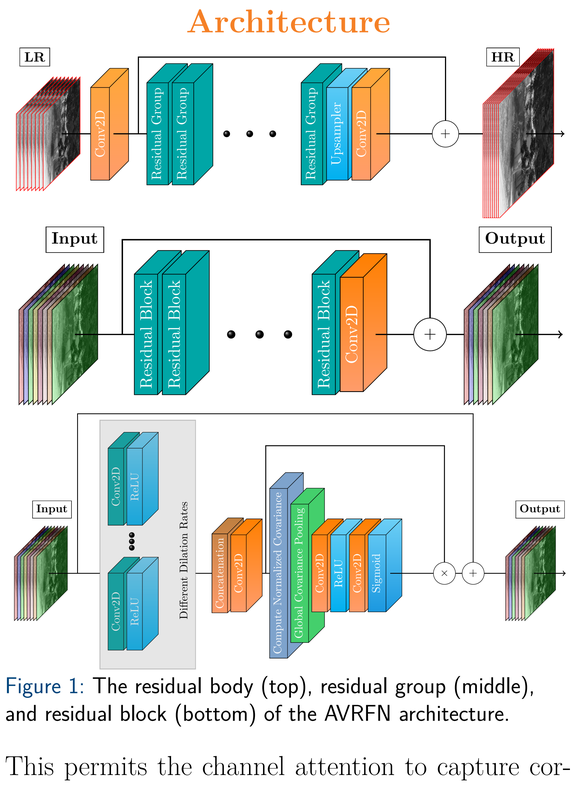

Figure 1: The residual body (top), residual group (middle), and residual block (bottom) of the AVRFN architecture.

This permits the channel attention to capture correlations higher than the first order.

Results

Model Variant	SOCA	Different Dilation Rates Module	Dilated Convolutions
DDRR		✓	✓
RRSOCA	✓		
CRCAN			✓
AVRFN	✓	✓	✓
RCAN			

Table 1: The model variants used during the ablation study.

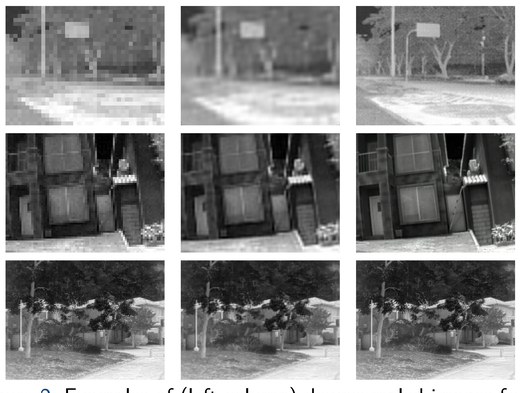

Figure 2: Examples of (left column) downsampled images from (top row) low-resolution, (middle row) medium-resolution, and (bottom row) high-resolution thermal cameras, their $\times 4$ upscaled counterparts (middle column), and the ground truth (GT) (right column).

Test Set	Model	Scale	Parameters	PSNR	SSIM
AXIS Domo P1290	RRSOCA	4	1661377	25.487	0.691
	DDRR	4	2839873	25.458	0.691
	CRCAN	4	2839873	**25.491**	**0.692**
	RCAN	4	1661377	25.239	0.682
	AVRFN	4	1917313	25.368	0.685
AXIS Q2901	RRSOCA	4	1661377	28.167	**0.802**
	DDRR	4	2839873	28.159	0.801
	CRCAN	4	2839873	**28.189**	0.802
	RCAN	4	1661377	27.923	0.795
	AVRFN	4	1917313	27.990	0.797
FLIR FC-6320	RRSOCA	4	1661377	31.978	0.867
	DDRR	4	2839873	31.985	0.867
	CRCAN	4	2839873	**32.002**	**0.867**
	RCAN	4	1661377	31.756	0.861
	AVRFN	4	1917313	31.824	0.864
TDAT	RRSOCA	4	1661377	28.388	0.641
	DDRR	4	2839873	**28.427**	**0.645**
	CRCAN	4	2839873	28.426	0.645
	RCAN	4	1661377	28.271	0.636
	AVRFN	4	1917313	28.298	0.637
KAIST	RRSOCA	4	1661377	**37.977**	**0.949**
	DDRR	4	2839873	37.456	0.918
	CRCAN	4	2839873	37.573	0.922
	RCAN	4	1661377	37.089	0.938
	AVRFN	4	1917313	37.827	0.943

Table 2: Ablation study results showing peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) scores on the Thermal Image Super-Resolution (TISR), FLIR Thermal Dataset for Algorithm Training (TDAT), and KAIST datasets for several model variants.

We conducted an ablation study to evaluate how CTD, our different dilation rates module, and SOCA interact in an SR network. In Table 2, we show the evaluation results for model variants displayed in Table 1. The qualitative results of our best performing model variant (CRCAN) are shown in Fig. 2. Each of our model variants performs better than the strong RCAN baseline. Additionally, SOCA shows clear performance gains. However, CRCAN, which only has dilated convolutions, attains the greatest performance.

Acknowledgements

The authors acknowledge TACC at the University of Texas at Austin for providing software, computational, and storage resources that have contributed to the research results reported within this project.

UNIVERSITY OF TEXAS ARLINGTON